口語文に含まれる固有名詞の属性自動識別手法の研究 : 関西弁コーパスを用いて

口語文に含まれる固有名詞の属性自動識別手法の研究

－関西弁コーパスを用いて－

岩井　拓也*

【要旨】

自然言語処理技術は検索エンジン、音声対話システムに伴い急速に発展し、医療[1]、教育[2]などでも使われるようになった。コンピュータが自然言語を処理するときは形態素解析が必須で、正常に処理するには、高い解析精度が必要となる。しかし、自然言語に口語表現や方言、未知語が含まれると、形態素解析の精度が低下することが、先行研究で指摘されている[3,4,5]。そこで本研究では、未知語を多く含む関西弁口語文からなる関西弁コーパスを用いて、形態素解析で生じる誤処理の自動修正、特に固有名詞の属性識別問題に取り組む。固有名詞は新語が次々に登場し、種類も多様で略語や造語など曖昧な表現が多いため、固有名詞の属性の一意な識別が難しいという特徴を持つ。本研究では関西弁コーパスに存在する固有名詞の属性を、適切に識別する手法を開発した。まず誤処理された形態素を自動収集し辞書登録をする手法を提案する。そしてこの識別で得られた固有名詞ごとの識別ラベル値を統合して、固有名詞の属性を機械学習で識別するためのデータを構築する。このデータに対して、機械学習の手法であるランダムフォレストにより固有名詞の識別を行なった結果、関西弁コーパスに含まれる 9 割近くの固有名詞の属性を、適切に識別することが可能となった。

キーワード：形態素解析、自然言語、関西弁コーパス、 固有名詞、ネット集合知

1. はじめに

近年、人間が日常的に用いる言語（自然言語）をコンピュータに処理させるための技術である自然言語処理が注目されている。自然言語処理は検索エンジン、音声対話システム、翻訳システムなど我々の日常に広く活用されている。さらにその適用分野は医療[1]、教育[2]などにも、拡がりつつある。また深層学習の発展で音声認識の精度や翻訳の精度が大きく向上し、研究分野としても注目されている。

形態素解析という基礎解析技術は、現在の自然言語処理で欠かせない処理となっている。形態素解析とは、自然言語の文章を形態素解析器を用いて、意味を持つ最小単位の形態素に分割し、各形態素に対して品詞等のタグを付与する処理である。多くの自然言語処理はまず、与えられた文章を形態素に分解してから、解釈、抽出などの計算機処理を行う。そのため自然言語処理では、形態素への適切な分割が必須の技術となる。

* 関西学院大学大学院総合政策研究科博士課程前期課程（ele00169@kwansei.ac.jp）

しかし日本語の形態素解析では、英語のような空白が単語間にないため、形態素解析の際に誤った境界で分割する、誤処理が頻繁に起きる。特に口語文は、書き言葉に比べ、言い直しや繰り返しなどの「非流暢性」、発話そのものが文になっていない「非文法性」などの問題で、形態素解析の精度が低下することが、先行研究[3, 4]などで指摘されている。さらにこうした口語文に方言が含まれると、方言特有の語の表記や、活用形、連接規則により形態素解析の精度が低下する[5, 6, 7, 8]。したがって口語文で方言が含まれ、近しい人間関係で交わされる会話は、自然言語の中で特に計算機処理が難しいとされる。

本研究は、このような方言を含む会話文の自然言語を対象とする。そして形態素解析で誤処理が高い頻度で起きる固有名詞の処理、特に固有名詞の人名、地名といった属性識別に、焦点をあてる。固有名詞は、流行語や登場する有名人、人気者など、次々に新語が未知語として誕生する。さらに短縮形やニックネームなど同一の対象を指す多様な固有名詞がある。そこで固有名詞の識別手法としては、増加する固有名詞への対応ができるだけ少ないマンパワーで可能であること、そして識別精度の向上や、システム維持が容易であることが、識別性能以外に要請される。

2. 本研究の目的

本研究の目的は、くだけた関西弁の会話から構成された関西弁コーパスを対象に、形態素解析の精度向上に資する手法を開発することである。そのために未知語が多く含まれる固有名詞の属性識別手法に取り組む。ここで固有名詞の属性とは、形態素解析システム MeCab が出力する固有名詞の品詞細分類「一般」「地域」「組織」「人名」を指す。固有名詞には新語の登場への対応、多義的に使われる属性の正確な識別が必要となる。さらにその識別性能の向上を、できるだけ低いマンパワーで行う仕組みの組み込みが求められる。そこで本研究では、こうした仕組みを実現できる手法を提案、実装し、その性能評価を行うことを目的とする。

3. 本研究で使用したデータ

本研究では、関西弁コーパス[9]を元に以下に述べる 3 種類のデータを作成し、計 4 種類のデータを用いる。

1 つ目は、ヘファナン・ケビン教授より提供を受けた関西弁コーパス[9]である。このデータは、二者間の関西弁を含む会話文が録音されたデータをテキストデータ化し、形態素解析器 MeCab による形態素解析と関西弁コーパス独自の話者コードを付与した上で、形態素解析の誤りをヘファナン・ケビン教授が手作業で修正したデータである。本研究で利用した関西弁コーパスは全 186 ファイルから構成され、重複を含めた約 225 万の形態素と記号を含む。

2 つ目は関西弁コーパスの形態素から復元した原文データ「関西弁口語文データ」である。口語文の原文は、本研究で提案する手法の性能評価のために必要である。しかし提供された関西弁コーパスには含まれていないので、原文を復元して作成する。

3 つ目は関西弁口語文データを用いて作成した「修正前関西弁注釈付きコーパス（以後、「修正前コーパス」と呼ぶ。）」である。この修正前コーパスは、関西弁口語文データを MeCab で形態素解析したデータである。修正を一切していないので、修正前コーパスと呼ぶ。

4 つ目のデータは修正後関西弁注釈付きコーパス（以後、「修正後コーパス」と呼ぶ。）である。修正後コーパスは「関西弁コーパス」の一部の記号を取り除き、MeCab と同じ出力形式に合わせたデータである。修正後コーパスはヘファナン・ケビン教授によって修正が加えられたデータである。本研究では精度評価をする際の正解データとして用いる。修正前コーパスと修正後コーパスの差分が、ヘファナン教授が手作業で行った修正箇所である。

4. 辞書登録による形態素解析の精度向上手法

本研究では、まず辞書登録による形態素解析の精度を向上する手法を提案する。具体的には、修正前コーパスと修正後コーパスの差分から誤り箇所の形態素を自動収集し、辞書に登録して辞書を作成、そしてその辞書を用いて形態素解析を行う。本手法を実装する前に、まず誤って形態素解析されたエラーを、エラータイプ 1〜4 の 4 タイプに分類し、各エラータイプの数や特徴を調べた。

エラータイプ 1 は単独の形態素を複数の形態素に誤って分割したエラーで、42,920 個存在する。続いてエラータイプ 2 は複数の形態素を単独の形態素に誤って分割したエラーのパターンで、27,834 個存在する。エラータイプ 3 は、形態素の分割に誤りは無いが紐付けする品詞等のタグが誤ったエラーで、215,169 個存在する。エラータイプ 4 は、形態素の分割する位置が誤っているエラーで、9,014 個存在する。

これらエラータイプのうち、エラータイプ 1 に該当する単独の形態素を辞書に登録し、単独の形態素が複数の形態素に誤って分割されるエラーの削減を目指す。まずエラータイプ 1 の形態素を辞書に登録するため、修正前コーパスと修正後コーパスと比較してエラー箇所を自動収集する。次にエラータイプ 1 に該当する単独の形態素を抽出し、単独の形態素を登録した辞書を作成、そしてこの辞書で再度、修正前コーパスの形態素解析を行う。なお辞書登録数によるエラー数の変化を調べるため、42,920 個存在する形態素を 10%刻みで辞書に登録し、登録後のエラー数の変化を確認する。

結果として、エラータイプ 1 から 4 を含む全体的なエラー率は、エラータイプ 1 の辞書登録割合が 0%〜70%までは減少した。しかし、エラータイプ 1 の割合が 70%を超えるとエラータイプ 1 は減少したが、エラータイプ 2 が増加し、全体のエラー率は悪化した。従って、エラータイプ 1 とエラータイプ 2 は、トレードオフの関係があることがわかった。このトレードオフの関係を考慮し、最もエラー率が低下するように登録する形態素や

MeCab の分割に用いるコストを変更して新たに辞書を作成する改善手法を検討した。しかし、エラータイプ 1 には品詞、特に固有名詞が最も多く、その未知語問題がエラータイプ 1 の本質的課題だとわかった。ここで提案した辞書登録手法は、次々と登場する新しい固有名詞を常に何らかの方法で別に把握し、辞書の登録更新作業を常に行わなければならない。そうなると辞書更新の手間が、実用上のネックとなることが予想される。

5. ネット集合知を用いた固有名詞の属性識別手法

次に二つ目の手法として、ネット集合知を用いた固有名詞の属性識別に取り組む。具体的には、3 つの Web サービス（Wikipedia、Google マップ、Google 画像検索）をネット集合知として活用し、固有名詞の属性（一般、地域、組織、人名）を識別する手法を提案、実装し、その後性能評価を行う。各 Web サービスを用いる手法として、たとえば Wikipedia の転送機能を利用した曖昧性の回避や顔認識技術を組み合わせた画像識別など、各サービスの特徴を踏まえた識別手法を提案する。また Web サービスごとに、処理した固有名詞それぞれについて、識別ラベル値という識別過程で得られるデータを収集する。この識別ラベル値は後述する機械学習を用いた手法でも使用する。

以下に Wikipedia を例にして識別ラベル値を得る手順の概略を述べる。まず識別する固有名詞を Wikipedia 上で検索し、表示されたページからカテゴリを取得する。次にカテゴリの一部を抽出し、独自に作成した属性照合テーブルを用いて照合する（表 1）。そして照合結果から識別ラベル値を付与し、そのラベル値で属性を決定する。識別ラベル値は、固有名詞の各属性（一般、地域、組織、人名）を識別するためのラベル、記事が存在しない場合に用いる「未照合」ラベル、そして記事は存在するが属性照合テーブルで照合できなかった場合の「決定不可能」ラベルの計 6 種類を用いる。

表 1. 属性照合テーブル（抜粋）

カテゴリの一部	属性
人物	人名
選手	人名
楽曲	一般
建築物	一般
企業	組織
・・・	・・・

Wikipedia 以外の Web サービスを用いた識別も同様の手続きである。まず固有名詞を Web サービス上で検索、次に定めた識別条件の結果に応じて識別ラベル値を反映、そして識別ラベル値の結果から属性を識別する。

表 2 に、これらの Web サービスをそれぞれ別個に利用し、固有名詞の属性を識別した結果を示す。識別率は、式 1 で与えられる。表 2 から、Web サービスごとで属性の識別率に違いがあり、Web サービスごとに特定の属性に対して識別率が異なることが、明らか

になった。例えば、Wikipedia は意外なことに「組織」、「一般」の識別率が比較的高く、「人名」の識別率は低かった。一方、Google マップは「地域」の識別率が顕著に高いことがわかった。Google マップは「地域」と「組織」のみ対象とした識別手法を提案したため、「一般」と「人名」は識別率がないので NA として表記する。Google 画像検索の識別率は予想通り「人名」が顕著に高い、という結果が得られた。

このネット集合知を用いる手法は「関西学院大学」のように属性が 1 つに限られる用語では適切に識別できた。一方で「三田」のように、地域や人名など複数の属性を持つ固有名詞は、適切な識別が難しいことがわかった。これらの結果から、属性識別の精度を向上するには、単独の Web サービスではなく、複数の Web サービスから得た識別結果を総合できる識別手法が必要となることがわかった。

$$\text{識別率}=\frac{\text{当該属性と識別された固有名詞数}}{\text{当該属性を持つ固有名詞数}} \times 100 \qquad (1)$$

表 2. 識別率の Web サービス別比較

	一般	地域	組織	人名
Wikipedia	31%	26%	48%	10%
Google マップ	NA	61%	10%	NA
Google 画像検索	NA	NA	12%	62%

6. ランダムフォレストを用いた固有名詞の属性識別手法

ネット集合知を個別で用いると、それぞれの特徴に応じて、得意とする識別属性が異なる。そこで、各 Web サービスの識別過程で得られた全固有名詞の識別ラベル値をデータとして用い、ランダムフォレストで総合的に固有名詞の属性識別を行う手法を検討する。ランダムフォレストは機械学習の手法の一つで、分類や予測によく用いられる。本研究のランダムフォレストの識別精度評価では「accuracy」を用いた。accuracy は実際の属性と識別した属性が一致した数を、識別に用いた固有名詞数で除した値に 100 を乗じた、式 2 で得られる数値である。

$$\text{accuracy}=\frac{\text{実際の属性と予測した属性が一致した数}}{\text{識別に用いた固有名詞数}} \times 100 \qquad (2)$$

ランダムフォレストの実装では Python の機械学習ライブラリの scikit-learn を用いた。scikit-learn ではランダムフォレストを実装に 19 個のパラメータが設定でき、このパラメータの値次第で精度が変わる。ただ 19 個のパラメータから最適な組み合わせを見つける

のは計算時間から困難である。そこで本研究では、accuracy に大きく影響するパラメータを特定し、そのパラメータのみ変化させる方法で適切なパラメータ設定を行う。

ランダムフォレストで固有名詞の識別を行うために、次のようにデータを新たに構築する。まず、説明変数（特徴量）は、5 節で述べた 3 つの Web サービスごとに識別を行った固有名詞から得られた識別ラベル値とする。それらの固有名詞に対して、関西弁コーパスから正解となる属性（一般、地域、組織、人名）を抽出し、それらの属性を被説明変数とする。新たな未知語でも精度よく識別できるように、構築したデータから 8 割をランダムフォレストのパラメーターチューニングに利用し、残り 2 割を属性識別の accuracy の評価に利用する。

表 3 に実際の属性とモデルで識別した属性の関係を示したクロス集計表を示す。行に本来この固有名詞の持つ属性、列は本手続きによって識別された属性をとり、数値は固有名詞を示す。accuracy は表 3 に太字で示した数値の合計(4,894)を分子、識別に用いた固有名詞数(5,469)を分母とする比率で、89.49%であった。ランダムフォレストを用いて Web サービス単独より属性識別精度は向上した。

一方で、識別できなかった固有名詞として「近鉄」（一般・組織）や「谷」（人名・地域）など、文脈によって複数の属性と解釈できる固有名詞が多くを占めた。このようなケースは文脈情報を用いて、より適切に識別することが必要である。

表 3. 実際の属性と識別した属性のクロス集計表

		識別した属性				合計
		一般	地域	組織	人名	
実際の属性	一般	**574**	39	22	38	673
	地域	112	**2748**	13	40	2913
	組織	93	8	**394**	13	508
	人名	115	66	16	**1178**	1375
合計		894	2861	445	1269	5469

7. 結論

本研究では形態素解析での誤処理の自動処理として、誤処理の中で多数を占める名詞、特に固有名詞の属性識別に取り組んだ。そして、3 つの手法を提案し、実装と性能評価を行った。

1 つ目の辞書登録による手法は、エラータイプ 1 とエラータイプ 2 がトレードオフの関係であること、そして次々と登場する新しい固有名詞を常に把握し、辞書に登録する更新作業が必要であることがわかった。2 つ目の手法では、次々と登場する新しい固有名詞に対処するため、3 つの Web サービスをネット集合知として使用し、固有名詞の属性識別を行った。この手法では、サービス単体ごとの得意とする属性と識別性能を明らかにした。

そして 3 つ目に、ネット集合知で得た識別データ値をまとめてデータとして使うランダムフォレストを用いた属性識別を実装し、評価を行った。

最終的にランダムフォレストで関西弁コーパスに含まれる 9 割近くの固有名詞の属性を、適切に識別することができた。これにより Wikipedia や Google 画像検索などの個別のサービスごと提案した識別ラベル値を求める手法が、機械学習の学習データの作成手法としても有効であることがわかった。

ランダムフォレストで適切に識別できた固有名詞には、人間でも属性識別が難しいと感じられる固有名詞が多く含まれていた。例えば、「九鬼」や「トノミネ[1]」を正しく地域と識別し、「小夜」「ボラム」「シヌ」を人名[2]と識別した。韓流アイドルの名前から局所的な地名まで、多岐にわたる固有名詞を一人の人間が網羅し、属性を識別することは一般的に難しい。従って、提案した手法は固有名詞の多様性、曖昧性に対処しながら、様々な固有名詞の属性識別を行う手法として有効といえる。またネット集合知を用いるため、次々と登場する新しい固有名詞に持続的に低い負荷で対応することができる。こうした意味でも提案手法は望ましいといえる。

ただし適切に識別できなかった固有名詞も残った。その多くが複数属性を持つ用語や、本研究で用いたネット集合知でも未照合になる用語（身内のみで使用される用語：大学のサークル名など）であった。従って今後の課題は、文脈を考慮した識別手法の考案、さらなるネット集合知への拡張とその識別手法の考案である。

【参考文献】

[1] 池田 伸,”自然言語処理による精神科入院患者の自殺リスク評価”，2020 年度 人工知能学会全国大会（第 34 回）論文集，1L3GS1305, 2020.

[2] 在間 拓幹，山本 利一，中村 茉耶，”人工知能の自然言語処理を利用したチャットボットを題材とした中学校技術科「双方向性のあるコンテンツのプログラミング」の授業実践”，教育情報研究，Vol. 35, No. 3, pp. 45-53, 2020.

[3] 松本 裕治，伝 康晴，”話し言葉の形態素解析”，情報処理学会研究報告自然言語処理，Vol. 2001, No. 54, pp. 49-54, 2001.

[4] 勝木 健太，笹野 遼平，河原 大輔，黒橋 禎夫，”Web 上の多彩な言語表現バリエーションに対応した頑健な形態素解析”，言語処理学会第 17 回年次大会発表論文集，pp. 1003-1006, 2011.

[5] 久留間 嵩之，乙武 北斗，吉村 賢治,”方言に起因する形態素解析の区切り誤りを自動検出する手法の試作”，電気関係学会九州支部連合大会講演論文集，Vol. 2016, p. 291, 2016.

[6] 廣川 純也，深澤 拓海，松村 冬子，原田 実，”形態素解析における関西弁の自動認識”, 研究報告自然言語処理，Vol. 2016, No. 3, pp. 1-7, 2016.

[7] 小林 聖也，奥村 紀之，”方言と標準語の違いを考慮した言語認識システムの開発”，2009 年度 人工知能学会全国大会（第 23 回）論文集，3I13, 2009.

[1] トノミネは砥峰高原を示す。
[2] ボラムは韓国の女優、シヌは韓国のアイドルグループメンバー

[8] 乙武 北斗，折舘 直樹，吉村 賢，”地方議会会議録の方言を含む発言における形態素解析誤りの分析”，日本知能情報ファジィ学会 第 30 回ファジィシステムシンポジウム講演論文集，pp. 660-663, 2014.

[9] ヘファナン・ケビン，”関西弁コーパスの紹介”，総合政策研究，No. 41, pp. 157-163, 2012.